International Journal of Legal Medicine (2021) 135:665–675
https://doi.org/10.1007/s00414-020-02489-5

ORIGINAL ARTICLE

Comparison of different machine learning approaches to predict dental age using Demirjian's staging approach

Antoine Galibourg[1,2] · Sylvain Cussat-Blanc[3,4,5] · Jean Dumoncel[2] · Norbert Telmon[1,2] · Paul Monsarrat[1,5,6] · Delphine Maret[1,2]

Received: 28 July 2020 / Accepted: 9 December 2020 / Published online: 7 January 2021

Abstract

Context Dental age, one of the indicators of biological age, is inferred by radiological methods. Two of the most commonly used methods are using Demirjian's radiographic stages of permanent teeth excluding the third molar (Demirjian's and Willems' method). The major drawbacks of these methods are that they are based on population-specific conversion tables and may tend to over- or underestimate dental age in other populations. Machine learning (ML) methods make it possible to create complex data schemas more simply while keeping the same annotation system. The objectives of this study are to compare (1) the capacity of ten machine learning algorithms to predict dental age in children using the seven left permanent mandibular teeth compared to reference methods and (2) the capacity of ten machine learning algorithms to predict dental age from childhood to young adulthood using the seven left permanent mandibular teeth and the four third molars.

Methods Using a large radiological database of 3605 orthopantomograms (1734 females and 1871 males) of healthy French patients aged between 2 and 24 years, seven left permanent mandibular teeth and the 4 third molars were assessed using Demirjian's stages. Dental age estimation was then performed using Demirjian's reference method and various ML regression methods. Two analyses were performed: with the 7 left mandibular teeth without third molars for the under 16 age group and with the third molars for the entire study population. The different methods were compared using mean error, mean absolute error, root mean square error as metrics, and the Bland-Altman graph.

Results All ML methods had a mean absolute error (MAE) under 0.811 years. With Demirjian's and Willems' methods, the MAE was 1.107 and 0.927 years, respectively. Except for the Bayesian ridge regression that gives poorer accuracy, there was no statistical difference between all ML tested.

Conclusion Compared to the two reference methods, all the ML methods based on the maturation stages defined by Demirjian were more accurate in estimating dental age. These results support the use of ML algorithms instead of using standard population tables.

Keywords Dental age · Machine learning · Artificial intelligence · Demirjian

Introduction

Estimated age is an important characteristic of an individual's biological identity and should be differentiated from the chronological age. In a living individual, chronological age generally corresponds to calendar age, whereas the biological age or physiological age covers the growth period, which corresponds to the maturation process of

Paul Monsarrat and Delphine Maret contributed equally to this work.

✉ Antoine Galibourg
antoinegalibourg@gmail.com

[1] Dental Faculty, Toulouse Institute of Oral Medicine, Paul Sabatier University, Toulouse University Hospital (Hôpitaux de Toulouse), chemin des Maraîchers, 31062 Toulouse Cedex 09, France

[2] Laboratoire Anthropologie Moléculaire et Imagerie de Synthèse, UMR 5288, Université Paul Sabatier, Toulouse, France

[3] Institute of Advanced Technologies in Living Sciences (ITAV), CNRS – USR3505, Toulouse, France

[4] Institute of Research in Informatics (IRIT) of Toulouse, CNRS – UMR5505, University of Toulouse, Toulouse, France

[5] Artificial and Natural Intelligence Toulouse Institute ANITI, Toulouse, France

[6] STROMALab, CNRS ERL 5311, EFS, ENVT, Inserm, UPS, Université de Toulouse, Toulouse, France

different tissues and organs [1], and the aging phase of an individual [2].

Dental age is one of the indicators of maturity, like skeletal age, or changes in the secondary sexual or somatic characteristics. Dental age is used in clinical practice in orthodontics, pedodontics, pediatrics, and orthopedic surgery. It is also widely used in physical anthropology and forensic science, for the identification of victims of mass disasters, to solve criminal, civil, or administrative issues such as adoption, migration, child abuse, legal consent, asylum proceedings, marriage, social benefits, and age estimation in competitive sports [3–7].

In living individuals, dental age estimation based on dental development is applicable during the period from childhood to young adulthood, that is to say, over a period that extends until all teeth have completed their maturation process, which corresponds to a maximum dental age based on maturation. The estimation methods must therefore be as reliable (range and percentage of the confidence limits) and accurate (mean difference between dental and chronological age) as possible [3, 8–10].

Radiological methods for assessing dental age are based on the continuous process of dental mineralization and allow radio-morphological (i.e., Demirjian's or atlas strategies) or radiometric (i.e., technical camera) observations [11, 12]. The number of tooth development stages described varies from 3 to 27 [13]. The challenge, therefore, lies in choosing a method that has enough stages to follow the continuous growth process as closely as possible and stages that are sufficiently differentiated to be easily discriminated by investigators [14]. Demirjian's dental maturity method was the first to combine outline drawings, written descriptive criteria, radiographic examples of each stage, and rules of choice to decide on borderline stages [13]. Examples where dental age is either early or late in relation to chronological age have been documented when this method is applied to other populations [10]. Population sample (regional background, sexual dimorphism, chronological age distribution), cultural or environmental factors (e.g., socio-economic status, nutrition, dietary habits), and also endocrine factors, or even reasons of statistical methodology [3] may explain such differences. The dental maturity scoring system is nevertheless accepted worldwide for its ease of use and for its excellent inter- and intra-examiner consistency [10, 15]. Beside the initial use of only 7 permanent teeth inside the statistical model (first mandibular incisor to second mandibular molar), the methodology was improved by adding other indicators, such as the third molars, to increase the identifiable age range up to young adults [16–19]. Whether the aim is to infer chronological age or to position an individual in relation to an age threshold, population-specific standards rather than a universal standard or methods developed on other populations need to be used [10]. At the present time, many studies assess the accuracy of a reference method on a population different from that used in the original method [18, 20–32]. Although these studies may reveal overestimation or underestimation of a method carried out on a sample of the reference population, they do not give a tool that is directly usable in forensic science to answer the questions of age estimation or classification. Depending on the populations studied, the DEM method has a tendency to overestimate dental age and can give average differences between chronological age and dental age ranging from −2.82 to +0.10 years for females and from −3.04 to +0.23 years for males [33]. Compared to the DEM method, the WILL method overestimates to a lesser extent the age of the children [34].

Machine learning (ML) algorithms are part of artificial intelligence that may represent a powerful tool to develop more accurate algorithmic processes to estimate an individual's dental age (regression problem) or define an age range (classification problem). ML is useful to discover patterns on large scale and multidimensional datasets that would have been difficult for individuals, even well trained, to identify [35]. Machine learning corresponds to a wide range of algorithm families and complexities, from easily interpretable models (e.g., linear regression, polynomial regression) to other models requiring development of additional explainability pipelines (e.g., support vector machine, neural network). The models can be supervised (the train dataset is "labeled") or unsupervised (the model works on its own to discover information, dealing with unlabeled data). These techniques therefore make it possible to provide more specific tools for each population and to determine more accurately the equivalent of the standards table for the population [3]. Machine learning methods can be parametrized at several levels: the hyperparameters are parameters that are determined a priori for each model, whereas the parameters are determined a posteriori, after the training stage. If a forensic utilization is envisaged, these parameters must be identified to comply with Ritz-Timme et al. recommendations for age estimation [36]. ML methods are already used for bone age estimation [37] and are starting to develop for dental age [38–40]. ML methods provide a solution to formulate models capable of capturing complexity without the need for correspondence tables while preserving the same annotation system. ML methods are described as being more accurate than conventional radiological methods [38, 40]. These methods provide mean absolute error estimates of dental age that are less than 0.99 years relative to chronological age [38–40]. The advantages of these methods are the absence of the use of conversion tables specific to a given population, the adaptability of the methods for each population studied, the possibility of directly giving an estimate of the dental age for use in forensic science, the possibility of integrating more predictors than those used in the classical methods, and thus of widening the range of estimation of the dental age that can be estimated from children to the young adult population. Although these methods have been evaluated in comparison with classical reference methods [38, 40], no comparison between the use of ML methods on different populations has yet been made. In this work, supervised ML methods have been used, since the statistical models have been created through optimization of the model deduced from the "known" dataset [38].

The objectives of this study are to compare (1) the capacity of ten machine learning algorithms to predict dental age in children using the seven left permanent mandibular teeth compared to reference methods and (2) the capacity of ten machine learning algorithms to predict dental age from childhood to young adulthood using the seven left permanent mandibular teeth and the four third molars.

Materials and methods

Data collection

A sample of 3605 panoramic radiographs (age ranging from 2.18 to 24 years) of 1734 girls and 1871 boys was retrospectively collected from the clinical radiological database of the Department of Oral Medicine of the Toulouse University Hospital, corresponding to a large sample of the population of Southern France population irrespective of ethnicity. Radiographs of individuals with systemic diseases or developmental anomalies were excluded. Data were treated as cross-sectional, although a proportion of children from each group was represented by more than one radiograph. According to the guidelines provided by Schmeling et al. [9], we have been working to get an even distribution of subjects across all age groups, a balanced number of boys and girls, and a minimum number of subjects 10 times the number of examined features (7 permanent teeth, 4 third molars and gender, i.e., 12 features examined in this study). As a minimum of 10 boys or girls in each category per year of age is required, the 15 individuals of 3 years old of the sample have been excluded [13]. Since the birth date and the date of the radiological examination are known for each subject, the chronological age is calculated as the number of years elapsed between these two dates and recorded as two decimal points.

Ethical considerations

According to the French health regulations and law governing ethics in human research (Public Health Code), retrospective studies based on the exploitation of usual care data do not need to be submitted to an ethics committee but have to be declared or covered by reference methodology of the French National Commission for Informatics and Liberties (CNIL). A collection and computer processing of personal and medical data was implemented to analyze the results of this research. Toulouse University Hospital signed a commitment of compliance to the reference methodology MR-004 of the CNIL (number: 2206723 v0).

Radiograph analysis

During the assessment of all panoramic radiographs, the observers were blinded for the chronological age. The developmental stages of each of the seven left permanent teeth of the mandible, from central incisor to the second molar, were defined according to the eight radiographic stages defined by Demirjian et al. [11]. Tooth formation stages from A to D represent maturation stages of the crown, and stages from E to H represent maturation stages of the roots. When one or more index teeth were absent, its contralateral homologue was staged. The developmental stages of the third molars were assessed according to the same method by which an extra stage was added for the radiolucent bud (stage 1) [41]. Then each alphabetical stage was converted into a numerical development score (stage 1, 1; stage A, 2; stage B, 3; stage C, 4; stage D, 5; stage E, 6; stage F, 7; stage G, 8; and stage H, 9). When a tooth or its contralateral was absent, the developmental stage was set to zero (absence, 0). The conditions for estimating dental age by Demirjian's method are only possible for individuals aged between 3 and 16 years. For this reason, a subgroup was identified for individuals under 16 years of age (U16 group) of the total sample (U24 group). To belong to the U16 group, the 7 mandibular teeth must have a maturity score. To belong to group U24, the stages of maturation of the 7 permanent teeth and all the third molars were considered. In this case, at least 4 permanent teeth from central incisor to second molar had to be staged, but all third molars could be absent.

Statistical analysis

Three months after the first assessment of all samples by the first observer (AG, confirmed and experienced dental practitioner), intra-observer reliability was tested by re-examining 50 panoramic radiographs randomly selected from the U24 group. Inter-observer agreement was tested with a second observer (LG, dental hospital intern) by assessing 50 randomly selected panoramic radiographs. The agreements were quantified using weighted Kappa statistics.

Two reference methods were used for dental age estimation on the U16 group: the one described by Demirjian et al. [11] and the one described by Willems et al. [19]. Demirjian's method (DEM) to infer dental age is a three-step process. Firstly, each stage of the seven permanent teeth is converted into seven numerical biologic weights with the help of the sex-specific conversion table. Secondly, the weights are added together to give a dental maturity score. Thirdly, sex-specific tables of standards are used to convert the dental maturity score into dental age. Willems' method (WILL) is easier because all numerical biologic weights of the seven permanent teeth were added together to directly give the dental age.

To validate the different ML strategies compared to the reference methods, the U16 and U24 datasets were randomly divided into a training dataset and test dataset in an 80–20% proportion, respectively. Gender was also used as a predictive feature. The ML approaches used training on two sources of information: the gender and the stages allocated by the main observer to the seven lower left permanent teeth and all third molars. The target values were the chronological ages. The following ML supervised regression algorithms were tested [42]: random forest (RF), support vector machine (SVM), decision tree (DT), Bayesian ridge regression (BRR), k-nearest neighbors (KNN), boosting method AdaBoost (ADAB), polynomial regression (POLYREG), and multi-layer perceptron (MLP). A pipeline was implemented for each ML algorithm, with the possibility of adding a variable normalization technique (min-max, standard scaling, or nothing) and a data reduction technique (principal component analysis or nothing). To test the benefit of combining the different techniques of ML mentioned above, stacking (STACK) [43] and voting (VOTE) [44] methods were also tested. The tuning of hyperparameters to obtain the best model was achieved by exploring multiple combinations using GridSearchCV function and a ten-fold cross-validation (to avoid overfitting, ensuring a 20% validation dataset to be used during hyperparameters optimization). The hyperparameters described in Supplementary Table 1 were tuned. To allow comparisons of the different machine learning methods, the entire procedure, starting with the random creation of new training and test sets, was repeated 20 times.

To allow ML methods to be compared with a reference method for individuals between 16 and 24 years of age, the age assessment based on the dental maturity score from the 4 third molars described by Hofmann et al. on a Central European population was carried out [45].

The accuracy of age predictions was measured by five indicators: the coefficient of determination (R^2), mean error (ME; chronological age minus predicted age), root mean square error (RMSE), mean square error (MSE), and mean absolute error (MAE). One-way ANOVA was performed to compare the different indicators across the different ML algorithms, and the multiple pairwise comparisons were adjusted using Tukey's HSD test [46].

The Bland and Altman method [47] was used to graphically assess the degree of agreement between chronological and biological age. All experiments were performed using Scikit-Learn 0.22.1 libraries [42] and Python 3.7.3. The significance level was set to .05.

Results

A total of 3570 orthopantomograms (1719 girls and 1851 boys) were finally considered for the group U24, from which 2230 orthopantomograms (1111 girls and 1119 boys) were in the subgroup U16. Age and sex distribution of the datasets are plotted in Fig. 1. The intra- and inter-observer weighted Kappa were 0.96 and 0.92, respectively.

Overall age prediction performance in the U16 age group

The age prediction performances are presented in Table 1. The mean error was obtained with chronological age minus predicted age; indeed, a negative value implies an

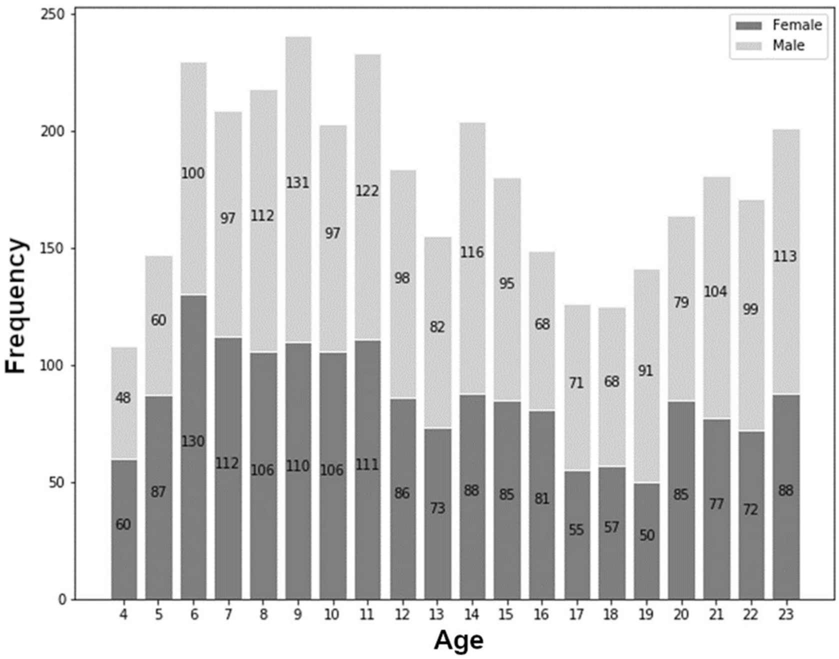


Fig. 1 Age and sex distribution for each category of age per year. *X*-axis represents age groups, 4 indicates all children from 4 to 4.99 years, etc.

Table 1 Mean error (ME), mean absolute error (MAE), mean square error (MSE), root mean square error (RMSE), and R^2 values for the U16 group (4–16 years old) assessing performance of machine learning regression methods and reference methods for chronological age estimation

Method	ME ± SD	MAE ± SD	MSE ± SD	RMSE ± SD	R^2 ± SD
DEM	− 0.705 ± 0.073 b (− 0.930 to − 0.582)	1.108 ± 0.045 c (1.029–1.256)	1.981 ± 0.175 b (1.734–2.572)	1.406 ± 0.060 c (1.317–1.604)	0.816 ± 0.018 b (0.758–0.839)
WILL	− 0.220 ± 0.075 c (− 0.460 to − 0.106)	0.928 ± 0.037 e (0.876–1.029)	1.418 ± 0.117 c (1.237–1.741)	1.190 ± 0.049 d (1.112–1.319)	0.868 ± 0.013 c (0.836–0.888)
BRR	− 0.002 ± 0.065 a (− 0.197–0.085)	0.812 ± 0.028 a (0.757–0.862)	1.030 ± 0.057 a (0.909–1.131)	1.014 ± 0.028 a (0.953–1.063)	0.904 ± 0.006 a (0.895–0.917)
SVM	0.016 ± 0.068 a (− 0.158–0.133)	0.729 ± 0.025 d (0.682–0.768)	0.901 ± 0.056 a (0.779–1.006)	0.949 ± 0.030 b (0.883–1.003)	0.916 ± 0.007 a (0.905–0.929)
DT	− 0.012 ± 0.064 a (− 0.174–0.106)	0.758 ± 0.034 b, d (0.692–0.858)	0.973 ± 0.089 a (0.814–1.255)	0.985 ± 0.044 a, b (0.902–1.120)	0.910 ± 0.010 a (0.882–0.927)
RF	− 0.007 ± 0.060 a (− 0.183–0.110)	0.731 ± 0.025 d (0.683–0.782)	0.885 ± 0.054 a (0.775–1.008)	0.940 ± 0.029 b (0.880–1.004)	0.918 ± 0.006 a (0.905–0.930)
KNN	0.009 ± 0.061 a (− 0.143–0.118)	0.738 ± 0.026 d (0.681–0.784)	0.921 ± 0.060 a (0.785–1.042)	0.959 ± 0.031 a, b (0.886–1.021)	0.915 ± 0.006 a (0.902–0.930)
MLP	− 0.041 ± 0.099 a (− 0.200–0.102)	0.742 ± 0.028 d (0.683–0.782)	0.907 ± 0.057 a (0.791–0.996)	0.952 ± 0.030 b (0.890–1.015)	0.916 ± 0.006 a (0.906–0.929)
POLYREG	− 0.008 ± 0.066 a (− 0.191–0.106)	0.735 ± 0.024 d (0.692–0.783)	0.913 ± 0.058 a (0.792–1.031)	0.955 ± 0.031 a, b (0.890–1.006)	0.915 ± 0.006 a (0.903–0.928)
ADAB	− 0.025 ± 0.064 a (− 0.225–0.075)	0.796 ± 0.026 a, b (0.755–0.846)	1.001 ± 0.058 a (0.886–1.099)	1.000 ± 0.029 a, b (0.941–1.048)	0.907 ± 0.006 a (0.896–0.920)
STACK	− 0.013 ± 0.066 a (− 0.191–0.074)	0.733 ± 0.025 d (0.684–0.783)	0.904 ± 0.052 a (0.791–1.012)	0.950 ± 0.027 b (0.890–1.006)	0.916 ± 0.006 a (0.905–0.929)
VOTE	0.068 ± 0.292 a (− 0.168–1.284)	0.770 ± 0.140 a, b, d (0.675–1.342)	0.995 ± 0.431 a (0.792–2.802)	0.984 ± 0.166 a, b (0.890–1.674)	0.908 ± 0.040 a (0.836–0.888)

For all metrics, the means ± standard deviations with (maximum-minimum) of the 20 independent experiments were given. Tukey's honest significance test for multiple comparison of means was performed at the .05 level. Within the same column, methods sharing the same letter are not significantly different from each other

Abbreviations: *ADAB* AdaBoost, *BRR* Bayesian ridge regression, *DT* decision tree, *KNN* k-nearest neighbor, *MLP* multi-layer perceptron, *POLYREG* polynomial regression, *RF* random forest, *STACK* stacking regressor, *SVM* support vector machine, *VOTE* voting regressor. Standard deviation (SD) of each method

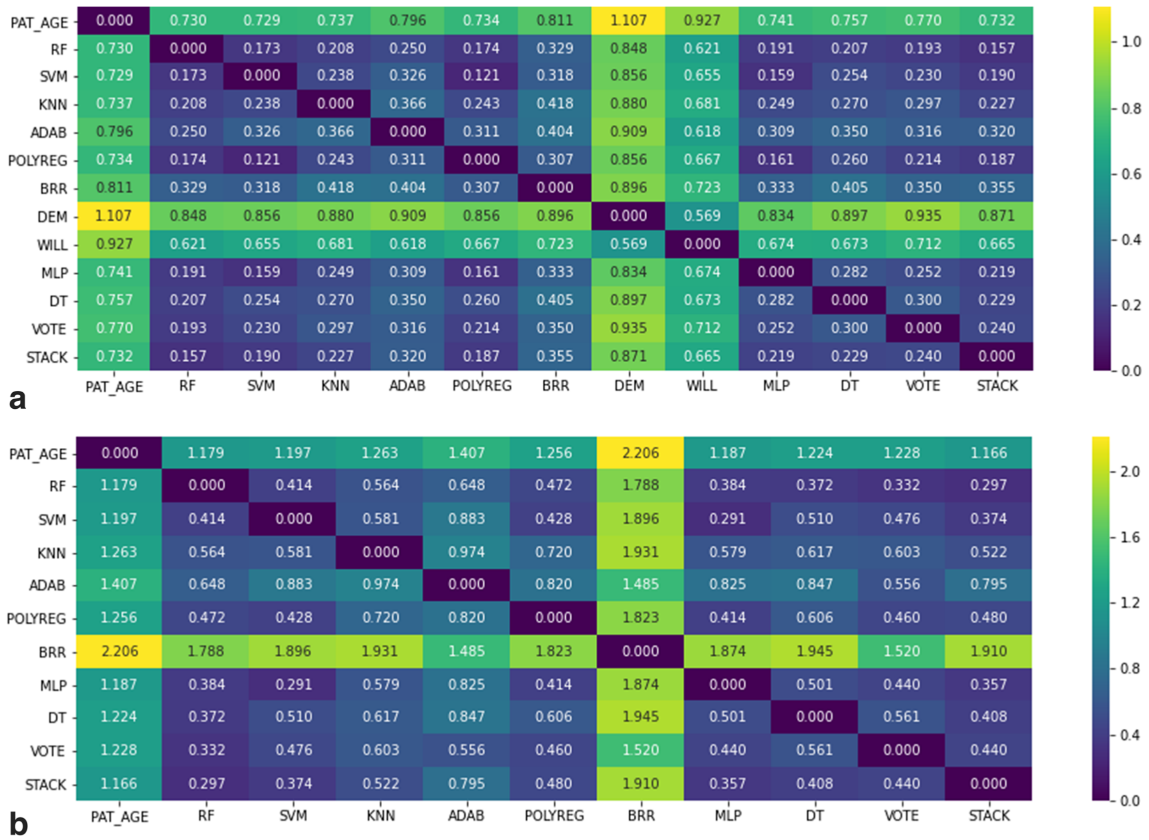
overestimation of the age estimation method compared to the real age. The two reference methods for estimating dental age, DEM, and WILL were used for U16. The DEM method significantly overestimates dental age more than the WILL method (− 0.71 ± 0.07 and − 0.22 ± 0.08, respectively). For all other metrics, the WILL method was more accurate than the DEM method.

All tested machine learning methods were significantly more accurate than the two reference methods for all metrics. ADAB and BRR achieved the lowest performance for MAE (Table 1, Fig. 2a).

Overall age prediction performance in the U24 age group

The age prediction performances are presented in Table 2 and Fig. 2b. The SVM and MLP methods result in a slight overestimation of dental age compared to chronological age (− 0.004 ± 0.071 and − 0.009 ± 0.148, respectively), while the other machine learning methods underestimate it. All metrics show significantly poorer performance for the ADAB and BRR methods. A non-significant trend shows better performance for the STACK, RF, MLP, and SVM methods.

Figure 3 a shows Bland and Altman graphs for the two reference methods DEM and WILL. An overestimation of dental age over the whole U16 group is observed for DEM (mean error − 0.70) and WILL (mean error − 0.22). The plots show a tendency of both methods to underestimate dental age up to the age of 6 years and to overestimate dental age from the age of 12 years. The outliers are similar for both methods. The onset of dental maturation is earlier for girls compared to boys. A dense cloud of dots is observed for both methods until the age of 8 years, and then alignments of dots parallel to each other are progressively identified more markedly with age. Each of the upward sloping lines corresponds to a combination of stages for the 7 left mandibular teeth. The spread of minimum and maximum chronological age for a combination of maturation stages is therefore visually observed with the extremities of these



	PAT_AGE	RF	SVM	KNN	ADAB	POLYREG	BRR	DEM	WILL	MLP	DT	VOTE	STACK
PAT_AGE	0.000	0.730	0.729	0.737	0.796	0.734	0.811	1.107	0.927	0.741	0.757	0.770	0.732
RF	0.730	0.000	0.173	0.208	0.250	0.174	0.329	0.848	0.621	0.191	0.207	0.193	0.157
SVM	0.729	0.173	0.000	0.238	0.326	0.121	0.318	0.856	0.655	0.159	0.254	0.230	0.190
KNN	0.737	0.208	0.238	0.000	0.366	0.243	0.418	0.880	0.681	0.249	0.270	0.297	0.227
ADAB	0.796	0.250	0.326	0.366	0.000	0.311	0.404	0.909	0.618	0.309	0.350	0.316	0.320
POLYREG	0.734	0.174	0.121	0.243	0.311	0.000	0.307	0.856	0.667	0.161	0.260	0.214	0.187
BRR	0.811	0.329	0.318	0.418	0.404	0.307	0.000	0.896	0.723	0.333	0.405	0.350	0.355
DEM	1.107	0.848	0.856	0.880	0.909	0.856	0.896	0.000	0.569	0.834	0.897	0.935	0.871
WILL	0.927	0.621	0.655	0.681	0.618	0.667	0.723	0.569	0.000	0.674	0.673	0.712	0.665
MLP	0.741	0.191	0.159	0.249	0.309	0.161	0.333	0.834	0.674	0.000	0.282	0.252	0.219
DT	0.757	0.207	0.254	0.270	0.350	0.260	0.405	0.897	0.673	0.282	0.000	0.300	0.229
VOTE	0.770	0.193	0.230	0.297	0.316	0.214	0.350	0.935	0.712	0.252	0.300	0.000	0.240
STACK	0.732	0.157	0.190	0.227	0.320	0.187	0.355	0.871	0.665	0.219	0.229	0.240	0.000

	PAT_AGE	RF	SVM	KNN	ADAB	POLYREG	BRR	MLP	DT	VOTE	STACK
PAT_AGE	0.000	1.179	1.197	1.263	1.407	1.256	2.206	1.187	1.224	1.228	1.166
RF	1.179	0.000	0.414	0.564	0.648	0.472	1.788	0.384	0.372	0.332	0.297
SVM	1.197	0.414	0.000	0.581	0.883	0.428	1.896	0.291	0.510	0.476	0.374
KNN	1.263	0.564	0.581	0.000	0.974	0.720	1.931	0.579	0.617	0.603	0.522
ADAB	1.407	0.648	0.883	0.974	0.000	0.820	1.485	0.825	0.847	0.556	0.795
POLYREG	1.256	0.472	0.428	0.720	0.820	0.000	1.823	0.414	0.606	0.460	0.480
BRR	2.206	1.788	1.896	1.931	1.485	1.823	0.000	1.874	1.945	1.520	1.910
MLP	1.187	0.384	0.291	0.579	0.825	0.414	1.874	0.000	0.501	0.440	0.357
DT	1.224	0.372	0.510	0.617	0.847	0.606	1.945	0.501	0.000	0.561	0.408
VOTE	1.228	0.332	0.476	0.603	0.556	0.460	1.520	0.440	0.561	0.000	0.440
STACK	1.166	0.297	0.374	0.522	0.795	0.480	1.910	0.357	0.408	0.440	0.000

Fig. 2 Heat maps showing the mean of the mean absolute errors (MAE) calculated from the 20 replicates for each pair of dental age estimation methods for the 4–16 years (**a**) and 4–24 years group (**b**). Abbreviations: ADAB, AdaBoost; BRR, Bayesian ridge regression; DEM, Demirjian's reference method; DT, decision tree; KNN, k-nearest neighbor; MLP, multi-layer perceptron; PAT_AGE, chronological age; POLYREG, polynomial regression; RF, random forest; STACK, stacking regressor; SVM, support vector machine; VOTE, voting regressor; WILL, Willems' reference method. Standard Deviation (SD) of each method

alignments. Both reference methods overestimate dental age for the more advanced combinations of tooth maturation and are more centered for the earlier stages.

Figure 3 b shows the results of the stacking learning machine method in the form of two Bland and Altman plots for group U16 and group U24 without and with the third molars, respectively. The dispersion of the measurements is greater for the U24 group from the age of 16 years, with maximal errors from 17 to 20 years, while decreasing after 20 years. Bland and Altman plots are available in supplementary data for all methods.

Discussion

Age estimation using the Demirjian and Willems reference methods

Compared to the world population

The first step in this work was to test our sample with two age estimation methods commonly referred in the literature. Among the reference methods, Demirjian et al. developed their method as a universal tool to assess dental maturity and predict dental age in children [11]. As expected from the existing literature, the results obtained by DEM method in this study overestimated the age of females by 0.18 years and males by 1.2 years. These results are consistent with the results of the meta-analysis of Jayaraman et al. which shows that the method overestimates the age on average of females by 0.65 years (− 0.10 to 2.82 years) and males by 0.60 years (− 0.23 to 3.04 years) [33]. Several authors have tried to overcome this problem, such as Willems et al. by distinct maturity score tables that give the age directly in years [19]. The results obtained by WILL method in this study underestimated the age of females by 0.36 years and overestimated the age of males by 0.77 years. These results are consistent with the results of the meta-analysis of Sehrawat et al. which shows that the method overestimates the age on average of females by 0.07 years (− 0.53 to 1.01 years) and males by 0.16 years (− 1.1 to 0.7 years) [34]. Although the meta-analysis of the results of these reference methods is consistent with our results, we can also compare them with a population that is geographically quite similar.

Table 2 Mean error (ME), mean absolute error (MAE), mean square error (MSE), root mean square error (RMSE), and R^2 values for the U24 group (4–24 years old), comparing machine learning regression methods for chronological age estimation

	ME ± SD	MAE ± SD	MSE ± SD	RMSE ± SD	R^2 ± SD
BRR	*0.017 ± 0.117 a* (− 0.217–0.228)	*2.206 ± 0.064 b* (2.094–2.358)	*7.822 ± 0.529 b* (6.932–9.202)	*2.795 ± 0.094 b* (2.633–3.034)	*0.759 ± 0.017 b* (0.731–0.789)
SVM	*− 0.004 ± 0.071 a* (− 0.135–0.152)	*1.197 ± 0.037 d, e* (1.129–1.258)	*2.758 ± 0.172 d, e* (2.512–3.042)	*1.660 ± 0.052 c, f* (1.585–1.744)	*0.915 ± 0.005 c, d* (0.906–0.928)
DT	*0.011 ± 0.071 a* (− 0.156–0.142)	*1.224 ± 0.042 c, d* (1.137–1.308)	*2.832 ± 0.216 c, d, e* (2.432–3.308)	*1.682 ± 0.064 c, d* (1.560–1.819)	*0.913 ± 0.006 c, d* (0.898–0.925)
RF	*0.012 ± 0.065 a* (− 0.115–0.144)	*1.179 ± 0.038 e* (1.113–1.258)	*2.595 ± 0.165 d* (2.329–3.024)	*1.610 ± 0.051 f* (1.526–1.728)	*0.920 ± 0.005 c* (0.906–0.928)
KNN	*0.044 ± 0.081 a* (0.087–0.224)	*1.263 ± 0.044 c* (1.201–1.352)	*3.087 ± 0.231 c* (1.667–1.898)	*1.756 ± 0.065 e* (1.667–1.898)	*0.905 ± 0.007 e* (0.886–0.913)
MLP	*− 0.009 ± 0.148 a* (− 0.279–0.234)	*1.187 ± 0.043 d, e* (1.127–1.275)	*2.701 ± 0.200 d* (1.568–1.778)	*1.643 ± 0.060 c, f* (1.568–1.778)	*0.917 ± 0.006 c* (0.902–0.926)
POLYREG	*0.003 ± 0.069 a* (− 0.135–0.152)	*1.256 ± 0.039 c* (1.176–1.309)	*2.995 ± 0.228 c, e* (2.586–3.437)	*1.729 ± 0.066 d, e* (1.608–1.854)	*0.908 ± 0.007 d, e* (0.894–0.920)
ADAB	*0.031 ± 0.082 a* (− 0.090–0.207)	*1.407 ± 0.049 a* (1.338–1.504)	*3.455 ± 0.212 a* (3.152–3.939)	*1.858 ± 0.057 a* (1.775–1.985)	*0.893 ± 0.006 a* (0.875–0.903)
STACK	*0.005 ± 0.064 a* (− 0.119–0.130)	*1.166 ± 0.037 e* (1.095–1.232)	*2.594 ± 0.175 d* (2.300–2.985)	*1.610 ± 0.054 f* (1.516–1.728)	*0.920 ± 0.005 c* (0.908–0.929)
VOTE	*0.014 ± 0.068 a* (− 0.093–0.140)	*1.228 ± 0.041 c, d* (1.167–1.296)	*2.770 ± 0.189 d, e* (2.538–max.233)	*1.663 ± 0.056 c, f* (1.593–1.798)	*0.915 ± 0.006 c, d* (0.900–0.923)

For all metrics, the means ± standard deviations with (maximum-minimum) of the 20 experiments were given. Tukey's honest significance test for multiple comparison of means was performed at the .05 level. Within the same column, methods sharing the same letter are not significantly different from each other

Abbreviations: *ADAB* AdaBoost, *BRR* Bayesian ridge regression, *DT* decision tree, *KNN* k-nearest neighbor, *MLP* multi-layer perceptron, *POLYREG* polynomial regression, *RF* random forest, *STACK* stacking regressor, *SVM* support vector machine, *VOTE* voting regressor. Standard deviation (SD) of each method

Compared to the population of Southwest France

The results obtained on a sample of population in the South of France with the two reference methods (i.e., Demirjian and Willems) are also in agreement with a previous study by Urzel and Bruznek, who used on a similar population, the same WILL age estimation method, and a variant of the DEM method revised by Demirjian and Goldstein [17, 24]: both methods overestimate dental age, and WILL method is more accurate than DEM method. For the WILL method, the results found by Urzel and Bruznek are more accurate, with a slight underestimation by 0.09 years for females and an overestimation by 0.14 years for males. These results can be explained by an age group distribution in this study with more individuals in the age groups below 7 years and above 13 years, namely, the age categories with the higher age differences in our study.

Age estimation using the machine learning methods

The second step of present work was to develop new standards based on machine learning strategies that can improve on the accuracy of actual table standards. Instead of developing a fixed standards table, machine learning may perceive complex data schemas. These strategies consider the workflow of ML methods (several algorithms may be chained) as well as the best hyperparameters determined a priori during training of the model.

In the literature, three main studies have used machine learning to estimate dental age and are discussed in this section: (a) the study by Tao et al. [38] with a multi-layer perceptron, using the same annotation system for the stages of tooth maturation; (b) the study by Štepanovský et al. [39] with several machine learning methods (multiple linear regression, support vector machine, multi-layer perceptron, radial basis function neural network, k-nearest neighbors, and k-star); and c) the study by Villa-Blanco et al. [40] with a convolutional neural network.

Comparison with reference methods

The first part of the evaluation of the ML methods was to compare them with the reference methods DEM and WILL under the same conditions of use of the latter, i.e., for the U16 population without using the third molars. All ML methods for estimating dental age outperformed the DEM and WILL reference methods. Among all machine learning methods

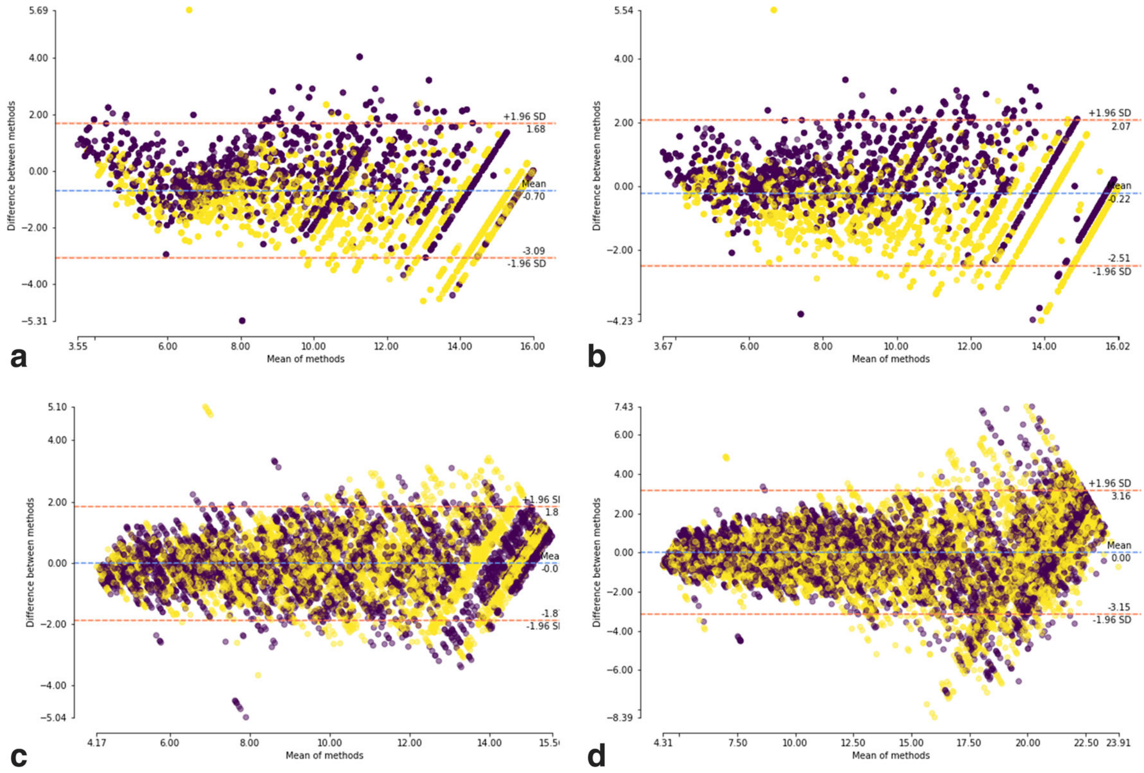


Fig. 3 Bland and Altman plot of difference in years between chronological age and dental age estimated **a** Demirjian's method (for the 4–16 years group), **b** Willems' method (for the 4–16 years group), **c** stacking regressor machine learning method (for the 4–16 years group), and **d** stacking regressor machine learning method (for the 4–24 years group). Negative values indicate that dental age is higher than chronological age. Purple points represent male individuals and yellow points represent female individuals. These graphs represent the 20 independent experiments performed

tested (4–16 years), the maximum mean differences obtained were an overestimation of 15 days for the multi-layer perceptron method and an underestimation of 25 days for the support vector machine, whereas DEM and WILL overestimated the age by a mean of 257 days and 80 days, respectively. These results are in agreement with the results of other studies on dental age estimation [38, 40].

The study by Metsäniitty et al. [18] assessed dental age in a U24 population using the 7 permanent mandibular teeth and the 4 third molars. The development of these teeth was staged respectively according to DEM method and Köhler et al. [48] The reported RMSE was 1.175 years for females and 1.095 years for males. These results are more accurate than those obtained for U24. This may be explained on the one hand by the third molar staging system using 10 stages of mineralization and on the other hand by the under-representation of individuals older than 17 years old in the sample.

Comparison with a study using a multi-layer perceptron

The study by Tao et al. [38] assessed one particular ML method that is a multi-layer perceptron, using the same annotation system for the stages of tooth maturation, and 7 permanent teeth. The MAE, MSE, and RMSE results of Tao et al. for MLP are 0.990, 1.775, and 1.332 years for males and 1.261, 2.616, and 1.617 years for females, compared to 0.742, 0.907, and 0.952 years for male/female in the present study in U16. These differences can be explained by the broader age boundaries used by Tao et al. For U24, using 7 permanent teeth + 4 third molars, the MLP MAE, MSE, and RMSE results are 1.187, 2.701, and 1.643 years.

Comparison with a study using a different annotation system

The study by Štepanovský et al. used several machine learning methods (multiple linear regression, support vector machine, multi-layer perceptron, radial basis function neural network, k-nearest neighbors, and k-star) [39]. However, Štepanovský used an annotation system described by Moorrees et al. employing more stages of mineralization of the dental organ [1]. The population sample of Štepanovský et al. [39] was 662 males and 314 females aged 2.7 to 20.5 years, with 16 teeth annotated on the left side of the maxilla and mandible on panoramic radiographs (14 stages of maturation). The MAE and RMSE results are under 0.94 and 1.27 years, respectively. The first reason for this better accuracy may be the frequency distribution of the population: 10% of the population was aged over 15 years compared to the 35% of the present study; the 15–20 age group exhibits the largest variations. The second possible reason is the use of more dental maturation stages and more teeth annotated than in our study: 14 versus 8 stages and 16 versus 11, respectively. Thus, the population of Štepanovský et al. may be compared to the U16 group of this study where the MAE and RMSE of all methods tested are under 0.811 and 1.014 years. Indeed, using a simpler annotation standard does not result in a significant performance loss compared to a higher number of teeth annotations.

Comparison with a study using a deep learning approach

The study by Villa-Blanco et al. used a convolutional neural network (CNN, a deep learning strategy) [40]. The authors succeeded in developing a fully automatic method of image analysis. Without using any annotation system, panoramic radiography is used straightforwardly. For the U16 group, the CNN obtained a ME and a MAE of − 0.020 ± 0.97 years and 0.75 years, respectively. For group U25, the CNN achieved a MAE of 1.21 ± 1.17 years. These results are similar with those obtained for U16 and U24, respectively.

The use of a CNN proposed by Vila-Blanco et al. is particularly attractive because it is fast, easy to use, and less subjected to human interpretation. It presents a disadvantage according the explainability of the results. Only areas of interest on the image are identified using a Grad-CAM methodology [49], whose heat map is mainly centered on the teeth. Nevertheless, the authors underline that Grad-CAM heat map remains "significantly variable according to the images."

Predictors used in ML methods

A strong advantage in forensic medicine relies on the combination of a proven annotation system (i.e., the maturation stages described by Demirjian) to create a statistical model adapted to the population to study. It was hypothesized that supervised machine learning is capable of interpreting data whose meaning is not accessible to us [35] and that the amount of data matters more than algorithms [50]. In the present study, sexual dimorphism (gender) has also been used as a predictor for ML, to avoid losing power by developing separate models for boys and girls.

Bland and Altman graphs to interpret the results of dental age estimation

The Bland and Altman graphs are informative at several levels. It can be visually observed that the reference methods tend to overestimate dental age for ages above 14 years. The ML methods show a more homogeneous distribution of points around the *x*-axis. We also observe for the same chronological age, the dispersion of the dental ages, which reflects both the individual variability of the maturation of the teeth, the imprecision of the machine learning methods, and the number of maturation stages. Thus, it is observed that even using third molars, the variability increases between 16 and 21 years of age (consequently the RMSE and MSE increase due to the square penalty). The Bland and Altman graph also allows us to observe the individual variability of a combination of maturation stages that corresponds to the alignment of points whose extremities correspond to individuals in the sample who are either early or late in development relative to the population sample. Finally, such graphs have an interest in the explicability of the ML methods used. Explicability of results is essential in forensic medicine. To be used in legal proceedings, it is necessary to provide elements of explanation. Some methods behave like black boxes (MLP, RF), while others are perfectly explainable (polynomial regression, decision tree). However, the methods the least easy to interpret give the best results on all the metrics (SVM, RF, MLP). New tools are being developed to break the black box and bring tangible elements of interpretation.

Drawbacks of machine learning methods tested

There are two approaches in forensic for the dental age. The first one estimates dental age and the second one aims at classifying individuals with respect to a key age [6]. A limitation of our study was to compare our results with those of the literature using reference methods to estimate dental age for the 16–24 age groups which is rather studied using a classification method. Prieto et al. [51] provided results comparable to ours on a Spanish population aged 14–21 years. They observe a mean error of − 0.1 ± 1.3 and − 0.07 ± 1.22 years respectively for the left and right third molar. Our results extracted from U24 for the 16–24-year-old subgroup give higher mean errors (Supplementary Table 2). Estimating the dental age for individuals aged from 16 to 24 years from the regression equation proposed by Hofmann et al. [45] gives results comparable (RMSE = 2.061 years) with those obtained with ML methods (Supplementary Table 2).

The use of dental age estimation methods, including machine learning methods in a forensic context, must consider individual variability. This is particularly important since the stages of dental maturation were developed in 1973 on a Canadian population [11]. Whatever the statistical approach, the variability of the prediction of dental age represents individual variability. Adding geographic origin, socio-nutritional conditions and endocrine factors using ML algorithm may help to reduce such individual variability [3] and improve dental age estimation. There is a trend towards maximum prediction accuracy for ML algorithms. This limit in accuracy may be due to the limitations of the annotation system and/or to inter-individual variability. A solution would be to go beyond the stages or to use additional predictors. Deep learning methods could allow a more detailed analysis of the maturation stages by breaking down the inter- or intra-examiner reproducibility. These methods are already used with CNN to analyze an entire panoramic radiograph to determine dental/bone age [40].

Conclusion

Machine learning methods were used in this study to assess the predictability of a dental maturation staging system based on dental panoramics. On the population studied in U16, as a reference, WILL method was significantly more accurate than DEM, and all ML methods were more accurate than the best reference method. On the population studied in U24, all ML methods rendered similar results, except for one (BRR). This study confirms that ML methods outperform the classic approaches for age estimation based on radiographic dental staging, from childhood up to early adulthood.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00414-020-02489-5.

Code availability The code generated during and/or analyzed during the current study is available from the corresponding author on reasonable request.

Authors' contributions Antoine Galibourg: Conceptualization, software, investigation, and writing—original draft

Sylvain Cussat-Blanc: Methodology and software

Jean Dumoncel: Writing—review and editing, software

Norbert Telmon: Validation and supervision

Paul Monsarrat: Formal analysis, methodology, software, writing—review and editing

Delphine Maret: Writing—review and editing, methodology

Data availability The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

Compliance with ethical standards

Conflicts of interest The authors declare that they have no conflicts of interest.

Ethics approval Not applicable.

Consent to participate Not applicable.

Consent for publication Not applicable.

References

1. Moorrees CFA, Fanning EA, Hunt EE (1963) Age variation of formation stages for ten permanent teeth. J Dent Res 42:1490–1502. https://doi.org/10.1177/00220345630420062701
2. Jylhävä J, Pedersen NL, Hägg S (2017) Biological age predictors. EBioMedicine 21:29–36. https://doi.org/10.1016/j.ebiom.2017.03.046
3. Braga J, Heuze Y, Chabadel O, Sonan NK, Gueramy A (2005) Non-adult dental age assessment: correspondence analysis and linear regression versus Bayesian predictions. Int J Legal Med 119: 260–274. https://doi.org/10.1007/s00414-004-0494-8
4. Cunha E, Baccino E, Martrille L, Ramsthaler F, Prieto J, Schuliar Y, Lynnerup N, Cattaneo C (2009) The problem of aging human remains and living individuals: a review. Forensic Sci Int 193:1–13. https://doi.org/10.1016/j.forsciint.2009.09.008
5. Willems G (2001) A review of the most commonly used dental age estimation techniques. Journal of Forensic Odonto-Stomatology 19:9–17
6. Sironi E, Gallidabino M, Weyermann C et al (2018) Bayesian networks of age estimation and classification based on dental evidence: a study on the third molar mineralization. J Forensic Legal Med 55:23–32. https://doi.org/10.1016/j.jflm.2018.02.005
7. Timme M, Steinacker JM, Schmeling A (2017) Age estimation in competitive sports. Int J Legal Med 131:225–233. https://doi.org/10.1007/s00414-016-1456-7
8. Galić I, Vodanović M, Cameriere R, Nakaš E, Galić E, Selimović E, Brkić H (2011) Accuracy of Cameriere, Haavikko, and Willems radiographic methods on age estimation on Bosnian-Herzegovian children age groups 6-13. Int J Legal Med 125:315–321. https://doi.org/10.1007/s00414-010-0515-8
9. Schmeling A, Geserick G, Reisinger W, Olze A (2007) Age estimation. Forensic Sci Int 65:178–181. https://doi.org/10.1016/j.forsciint.2006.05.016
10. Esan TA, Yengopal V, Schepartz LA (2017) The Demirjian versus the Willems method for dental age estimation in different populations: a meta-analysis of published studies. PLoS One 12:1–23. https://doi.org/10.1371/journal.pone.0186682
11. Demirjian A, Goldstein H, Tanner JM (1973) A new system of dental age assessment. Hum Biol 45:211–227
12. Thevissen PW, Fieuws S, Willems G (2011) Third molar development: measurements versus scores as age predictor. Arch Oral Biol 56:1035–1040. https://doi.org/10.1016/j.archoralbio.2011.04.008
13. Liversidge HM (2012) The assessment and interpretation of Demirjian, Goldstein and Tanner's dental maturity. Ann Hum Biol 39:412–431
14. Lucas VS, Andiappan M, McDonald F, Roberts G (2016) Dental age estimation: a test of the reliability of correctly identifying a subject over 18 years of age using the gold standard of chronological age as the comparator. J Forensic Sci 61(5):1238–1243. https://doi.org/10.1111/1556-4029.13132
15. Jayaraman J, Roberts GJ, Wong HM, King NM (2018) Dental age estimation in southern Chinese population using panoramic radiographs: validation of three population specific reference datasets. BMC Med Imaging 18:4–11. https://doi.org/10.1186/s12880-018-0250-z
16. Chaillet N, Nyström M, Demirjian A (2005) Comparison of dental maturity in children of different ethnic origins: international maturity curves for clinicians. J Forensic Sci 50:1–11. https://doi.org/10.1520/jfs2005020
17. Demirjian A, Goldstein H (1976) New systems for dental maturity based on seven and four teeth. Ann Hum Biol 3:411–421. https://doi.org/10.1080/03014467600001671
18. Metsäniitty M, Waltimo-Sirén J, Ranta H, Fieuws S, Thevissen P (2019) Dental age estimation in Somali children and sub-adults combining permanent teeth and third molar development. Int J Legal Med 133:1207–1215. https://doi.org/10.1007/s00414-019-02053-w
19. Willems G, Van Olmen A, Spiessens B, Carels C (2001) Dental age estimation in Belgian children: Demirjian's technique revisited. J Forensic Sci 46:893–895
20. Alassiry A, Alshomrani K, Al Hasi S et al (2019) Dental age assessment of 3-15-year-old Saudi children and adolescents using Demirjian's method-a radiographic study. Clinical and Experimental Dental Research 5:336–342. https://doi.org/10.1002/cre2.186
21. Prasad H, Kala N (2019) Accuracy of two dental age estimation methods in the Indian population - a meta-analysis of published studies. Journal of Forensic Odontostomatology 3:2–11

22. Ranasinghe S, Perera J, Taylor JA, Tennakoon A, Pallewatte A, Jayasinghe R (2019) Dental age estimation using radiographs: towards the best method for Sri Lankan children. Forensic Sci Int 298:64–70. https://doi.org/10.1016/j.forsciint.2019.02.053
23. Yang Z, Geng K, Liu Y, Sun S, Wen D, Xiao J, Zheng Y, Cai J, Zha L, Liu Y (2019) Accuracy of the Demirjian and Willems methods of dental age estimation for children from central southern China. Int J Legal Med 133:593–601. https://doi.org/10.1007/s00414-018-1924-3
24. Urzel V, Bruzek J (2013) Dental age assessment in children: a comparison of four methods in a recent french population. Journal of Forensic Science 58:1341–1347. https://doi.org/10.1111/1556-4029.12221
25. Alqadi MA, Abuaffan AH (2019) Validity of the Demirjian and Fishman methods for predicting chronological age amongst Yemeni children. Sultan Qaboos University Medical Journal 19: 26–31. https://doi.org/10.18295/squmj.2019.19.01.006
26. Balgi P, Katge F, Singh S, Jayaraman J (2020) Comparative assessment to determine the accuracy of Demirjian and DARLInG methods of dental age estimation in western Indian children. Legal Medicine (Tokyo) 42:101629. https://doi.org/10.1016/j.legalmed.2019.101629
27. Hegde S, Patodia A, Shah K, Dixit U (2019) The applicability of the Demirjian, Willems and Chaillet standards to age estimation of 5-15 year old Indian children. The Journal of Forensic Odontostomatology 37:40–50
28. Kar May L, Mei Shian AY, Durward C, Jayaraman J (2020) A method of estimating age of undocumented children and young adults of different socioeconomic status in Cambodia. Heliyon 6: e03476. https://doi.org/10.1016/j.heliyon.2020.e03476
29. Kermani M, Tabatabaei Yazdi F, Abed Haghighi M (2019) Evaluation of the accuracy of Demirjian's method for estimating chronological age from dental age in Shiraz, Iran: using geometric morphometrics method. Clinical and Experimental Dental Research 5:191–198. https://doi.org/10.1002/cre2.169
30. Lan LM, Yang ZD, Sun SL et al (2019) Application of Demirjian's and Cameriere's method in dental age estimation of 8-16 year old adolescents from Hunan Han nationality. Fa Yi Xue Za Zhi. 35: 406–410. https://doi.org/10.12116/j.issn.1004-5619.2019.04.005
31. Mohanty I, Panda S, Dalai RP, Mohanty N. Predictive accuracy of Demirjian's, modified Demirjian's and India specific dental age estimation methods in Odisha (Eastern Indian) population. The Journal of Forensic Odonto-stomatology 37:32–39
32. Moness Ali AM, Ahmed WH, Khattab NM (2019) Applicability of Demirjian's method for dental age estimation in a group of Egyptian children. BDJ Open 5:2. https://doi.org/10.1038/s41405-019-0015-y
33. Jayaraman J, Wong HM, King NM, Roberts GJ (2013) The French–Canadian data set of Demirjian for dental age estimation: a systematic review and meta-analysis. J Forensic Legal Med 20: 373–381. https://doi.org/10.1016/j.jflm.2013.03.015
34. Sehrawat JS, Singh M (2017) Willems method of dental age estimation in children: a systematic review and meta-analysis. J Forensic Legal Med 52:122–129. https://doi.org/10.1016/j.jflm.2017.08.017
35. Fabris F, de Magalhães JP, Freitas AA (2017) A review of supervised machine learning applied to ageing research. Biogerontology. 18:171–188. https://doi.org/10.1007/s10522-017-9683-y
36. Ritz-Timme S, Cattaneo C, Collins MJ, Waite ER, Schütz HW, Kaatsch HJ, Borrman HIM (2000) Age estimation: the state of the art in relation to the specific demands of forensic practise. Int J Legal Med 113:129–136. https://doi.org/10.1007/s004140050283
37. Dallora AL, Anderberg P, Kvist O, Mendes E, Diaz Ruiz S, Sanmartin Berglund J (2019) Bone age assessment with various machine learning techniques: a systematic literature review and meta-analysis. PLoS One 14:1–22. https://doi.org/10.1371/journal.pone.0220242
38. Tao J, Wang J, Wang A et al (2020) Dental age estimation: a machine learning perspective. In: Hassanien a., Azar a., Gaber T., Bhatnagar R., F. Tolba M. (eds) The International Conference on Advanced Machine Learning Technologies and Applications (AMLTA2019). AMLTA 2019. Advances in Intelligent Systems and Computing, vol 921. Springer, Cham. pp. 722–733. https://doi.org/10.1007/978-3-030-14118-9_71
39. Štepanovský M, Ibrová A, Buk Z, Velemínská J (2017) Novel age estimation model based on development of permanent teeth compared with classical approach and other modern data mining methods. Forensic Sci Int 279:72–82. https://doi.org/10.1016/j.forsciint.2017.08.005
40. Vila-Blanco N, Carreira MJ, Varas-Quintana P, Balsa-Castro C, Tomas I (2020) Deep neural networks for chronological age estimation from OPG images. IEEE Trans Med Imaging 39:2374–2384. https://doi.org/10.1109/TMI.2020.2968765
41. Duangto P, Iamaroon A, Prasitwattanaseree S, Mahakkanukrauh P, Janhom A (2017) New models for age estimation and assessment of their accuracy using developing mandibular third molar teeth in a Thai population. Int J Legal Med 131:559–568. https://doi.org/10.1007/s00414-016-1467-4
42. Pedregosa F, Varoquaux G, Gramfort A et al (2011) Scikit-learn: machine learning in Python. J Mach Learn Res 12:2825–2830 https://www.jmlr.org/papers/v12/ (accessed 1 December 2020)
43. Wolpert DH (1992) Stacked generalization. Neural Netw 5(2):241–259. https://doi.org/10.1016/S0893-6080(05)80023-1
44. Mu X, Lu J, Watta P, Hassoun MH (2009) Weighted voting-based ensemble classifiers with application to human face recognition and voice recognition. In: 2009 International Joint Conference on Neural Networks [Internet]. IEEE 2168–2171. https://doi.org/10.1109/IJCNN.2009.5178708
45. Hofmann E, Robold M, Proff P, Kirschneck C (2017) Age assessment based on third molar mineralisation: an epidemiological-radiological study on a central-European population. J Orofac Orthop 78:97–111. https://doi.org/10.1007/s00056-016-0063-z
46. Tukey JW (1949) Comparing individual means in the analysis of variance. Biometrics. 5:99–114
47. Bland JM, Altman DG (1986) Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1:307–310
48. Köhler S, Schmelzte R, Loitz C, Püschel K (1994) Die entwicklung des weisheitszahnes als kriterium der lebensaltersbestimmung [development of wisdom teeth as a criterion of age determination]. Ann Anat 176:339–345
49. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D (2020) Grad-CAM: visual explanations from deep networks via gradient-based localization. Int J Comput Vis 128:336–359. https://doi.org/10.1007/s11263-019-01228-7
50. Halevy A, Norvig P, Pereira F (2009) The unreasonable effectiveness of data. IEEE Intell Syst 24:8–12. https://doi.org/10.1109/MIS.2009.36
51. Prieto JL, Barbería E, Ortega R, Magaña C (2005) Evaluation of chronological age based on third molar development in the Spanish population. Int J Legal Med 119:349–354. https://doi.org/10.1007/s00414-005-0530-3

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.